

# Question Answering in Telugu Language

*Project-I (CS4D002) Report submitted in partial fulfillment for  
the award of degree of*

**Bachelors and Masters of Technology**

*in*

**Computer Science and Engineering**

*by*

**Thadicherla Hrishith  
20CS02002**

Under the supervision of

**Dr. Abhik Jana**



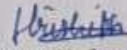
SCHOOL OF ELECTRICAL SCIENCES

INDIAN INSTITUTE OF TECHNOLOGY BHUBANESWAR

# Certificate

## Candidate's Declaration

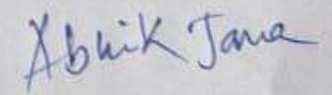
I hereby declare that the work presented in the project entitled "Question Answering For Telugu Language" in partial fulfillment of the requirements for the award of the Degree of Bachelors and Masters of Technology and submitted in the School of Electrical Sciences of the Indian Institute of Technology Bhubaneswar is an authentic record of my own work carried out under the supervision of Prof. Dr. Abhik Jana, School of Electrical Sciences, Indian Institute of Technology Bhubaneswar. The matter presented in this thesis has not been submitted by me for the award of any other degree of this or any other Institute/University.

  
Thadicherla Hrishith  
(Roll no. 20CS02002)

This is to certify that the above statement made by the candidate is true to the best of our knowledge and belief.

Place:  
Date:

IIIT BBSR  
03/05/24

  
Dr. Abhik Jana  
School of Electrical Sciences  
Indian Institute of Technology Bhubaneswar

# Acknowledgement

It is a great honor to express my most profound respect and sense of gratitude to my B. Tech Project supervisor **Dr. Abhik Jana** for his knowledge, insights, expertise, guidance, enthusiastic involvement, and persistent encouragement during the planning and development of this thesis work.

I also gratefully acknowledge his meticulous efforts in thoroughly going through and improving many of my research manuscripts without which this work could not have been completed.

I am highly obliged to all the professors of the Computer Science Department, including **Dr. Manoranjan Satpathy**, **Dr. D. P. Dogra**, **Dr. Joy Chandra Mukherjee**, **Dr. Srinivas Pinisetty**, and **Dr. Sudipta Saha** for providing all the guidance, help, and encouragement during my last four years at college.

I am extremely grateful to my parents and grandparents for their moral support, love, encouragement, and blessings to complete this task.

I would like to express my deep and sincere thanks to my friends and all other persons whose names do not appear here, for helping me either directly or indirectly in all even and odd times.

Finally, I am indebted and grateful to the Almighty for helping me in this endeavor.

# Abstract

Languages play an important role in communication whether it is among humans or machines. Around 6500 languages are spoken in the world, among those 1652 are from India. In the modern era of communication, everyone wants to get all the information in their native language to lead the ease of accomplishing their day-to-day life. Telugu is one of the widely spoken languages in southern parts of India. Except for some dialogue-based Question Answering(QA) systems, no other state-of-the-art model is implemented to do QA in web-based applications for the Telugu language.

The goal of this study is to try and enhance Machine Reading Comprehension (MRC) for low-resource languages like Telugu. Last semester, we worked on reproducing the results of two papers, one described by AVADHAN by Priyanka et al. [1](2020) for SVM Question Classifier and the other described by TeQuAD by Rakesh et al. [2](2022). This semester, we evaluate three additional models IndicBERT, XLM-Roberta and L3-Cube alongside previous approaches which involved using mBERT and compare what the results are and which model performs better. After evaluation, the XLM-RoBERTa model has an F1 score of 0.575, an exact match accuracy of 0.385, and a partial match accuracy of 0.736, L3-Cube model outperforms the others with an F1 score of 0.700, an exact match accuracy of 0.446, and a partial match accuracy of 0.603, Indic-BERT model has the lowest accuracies among the four, with an F1 score of 0.444, an exact match accuracy of 0.261, and a partial match accuracy of 0.351, m-BERT model has an F1 score of 0.594, an exact match accuracy of 0.356, and a partial match accuracy of 0.605. GitHub link to the codes and datasets is available at this [GitHub repository](#).

# Contents

<b>Acknowledgement</b>	<b>ii</b>
<b>Abstract</b>	<b>iii</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 Related Work</b>	<b>2</b>
<b>3 Background on Models</b>	<b>3</b>
3.1 IndicBERT . . . . .	3
3.2 XLM-RoBERTa . . . . .	3
3.3 L3-Cube . . . . .	4
<b>4 Methodology</b>	<b>5</b>
4.1 Dataset Description . . . . .	5
4.2 Data Preprocessing . . . . .	5
4.2.1 Loading the dataset . . . . .	5
4.2.2 Checking Context and Question Distributions . . . . .	5
4.2.3 Tokenization . . . . .	7
4.2.4 Appending Start and End Positions of Answer to Encoded data . .	7
4.3 Training the model . . . . .	9
<b>5 Results</b>	<b>11</b>
5.1 XLM-RoBERTa . . . . .	11
5.1.1 Results . . . . .	11
5.1.2 Cases where the model fails . . . . .	11
5.2 L3-Cube . . . . .	13
5.2.1 Results . . . . .	13
5.2.2 Cases where the model fails . . . . .	13
5.3 Indic-BERT . . . . .	14
5.3.1 Results . . . . .	14
5.3.2 Cases where the model fails . . . . .	14
5.4 m-BERT (Prior-Work) . . . . .	16

5.4.1	Results . . . . .	16
5.4.2	Cases where the model fails . . . . .	16
5.5	Cases where RoBERTa and L3-Cube give different results . . . . .	18
5.5.1	Case 1: Where RoBERTa fails and L3-Cube gives correct answer . .	18
5.5.2	Case 2: Where L3-Cube fails and RoBERTa gives correct answer . .	20
5.6	Results Summarised . . . . .	21
5.7	Conclusion . . . . .	21

# Chapter 1

## Introduction

Question-answering (QA) systems play a crucial role in the field of Natural Language Processing (NLP), particularly for low-resource languages like Telugu. These systems are designed to understand and respond to queries in a natural, human-like manner, making information more accessible and interactive.

The importance of QA systems for low-resource languages is multi-fold. Firstly, they help in preserving and promoting linguistic diversity. In an increasingly globalized world, there's a risk that low-resource languages could become marginalized or even extinct. By developing QA systems for these languages, we can ensure that they continue to thrive in the digital age.

Secondly, QA systems can make a significant impact in regions where low-resource languages are spoken. For instance, they can be used in educational settings to support learning, or in healthcare to provide information and advice. This can be particularly beneficial in areas where access to such resources might be limited.

However, developing QA systems for low-resource languages presents unique challenges. These languages often lack the large-scale labelled datasets that are available for languages like English. Therefore, innovative approaches are needed to overcome these challenges, such as leveraging semantic networks, augmenting question information, or using pre-trained language models.

# Chapter 2

## Related Work

While substantial progress has been made in English question-answering (QA) tasks with datasets like SQuAD [6] and CNN/Dailymail Chen et al. [7], the exploration of QA in other languages faces challenges due to the scarcity of annotated datasets. Some efforts have been made in this regard; for instance, Clark et al. [8] (2020) introduced a multilingual QA dataset covering 11 diverse languages, including Telugu.

Addressing the issue of low-resource languages, Hsu et al. [9] (2019) explored zero-shot cross-lingual transfer learning for reading comprehension tasks, suggesting that direct translation may not be necessary. Translation-based data augmentation, as proposed by Bornea et al. [10] (2020), emerged as a mechanism to enhance multilingual transfer learning.

To bridge the language gap, Liu et al. [11] (2020) and Cui et al. [12] (2019) proposed leveraging translated information from high-resource languages to perform well in low-resource languages. Cui et al. [13] (2019) introduced back-translation approaches and a novel model called 'Dual BERT,' aiming to enhance cross-lingual transfer learning by learning semantic information from bilingual QA pairs.

Yuan et al. [14] (2020) introduced phrase boundary supervision tasks to improve answer boundary detection in low-resource MRC models, trained on high-resource languages. Reddy et al. [15] (2020) addressed answer span improvement through post-correction methods, adding layers to transformer-based language models.

Exploring information retrieval systems, Zheng et al. [16] work on AnswerBus implemented a web-based QA system for multiple languages, including English and German. In the realm of Telugu QA, Bandyopadhyay et al. [17] dialogue-based QA architecture on railway information demonstrated an 83.96% dialogue success rate and 96.34% precision.

Additionally, DeepPavlov Mikhail et al. [18] Wikipedia-pretrained ODQA system, with components for ranking and reading, showcased promising results for both English and Russian Wikipedia articles.



# Chapter 3

## Background on Models

### 3.1 IndicBERT

IndicBERT is a multilingual ALBERT model pre-trained exclusively on 12 major Indian languages. It was introduced in Indic NLP Suite paper by Kakwani et al. [3]. It is pre-trained on our novel monolingual corpus of around 9 billion tokens and subsequently evaluated on a set of diverse tasks. IndicBERT has much fewer parameters than other multilingual models (mBERT, XLM-R etc.) while it also achieves a performance on-par or better than these models.

The 12 languages covered by IndicBERT are Assamese, Bengali, English, Gujarati, Hindi, Kannada, Malayalam, Marathi, Oriya, Punjabi, Tamil, and Telugu.

### 3.2 XLM-RoBERTa

XLM-RoBERTa is a multilingual version of RoBERTa. It is pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages including Telugu. It was introduced in the paper Unsupervised Cross-lingual Representation Learning at Scale by Conneau et al. [4]

RoBERTa is a transformers model pre-trained on a large corpus in a self-supervised fashion. This means it was pre-trained on the raw texts only, with no humans labelling them in any way (which is why it can use lots of publicly available data) with an automatic process to generate inputs and labels from those texts.

More precisely, it was pre-trained with the Masked language modelling (MLM) objective. Taking a sentence, the model randomly masks 15% of the words in the input then run the entire masked sentence through the model and has to predict the masked words. This is different from traditional recurrent neural networks (RNNs) that usually see the words one after the other, or from autoregressive models like GPT which internally mask the future tokens. It allows the model to learn a bidirectional representation of the sentence.

This way, the model learns an inner representation of 100 languages that can then be used to extract features useful for downstream tasks: if you have a dataset of labelled sentences for instance, you can train a standard classifier using the features produced by the XLM-RoBERTa model as inputs.

### 3.3 L3-Cube

L3Cube has developed a Hindi BERT model, known as L3Cube-HindBERT, which is pre-trained on a Hindi monolingual corpus. This model is designed to understand and process Hindi language text at a high level of sophistication. This is written by Raviraj et al. [5]

Interestingly, since Hindi and Marathi, two Indic languages, share the same script (Devanagari), L3Cube has trained a single model for both languages. This model, known as DevBERT, is trained on both Marathi and Hindi monolingual datasets.

These models have been evaluated on downstream tasks such as text classification and named entity recognition in both Hindi and Marathi. The results show that the HindBERT and DevBERT-based models significantly outperform other multi-lingual models like MuRIL, IndicBERT, and XLM-R.

Based on these promising results, L3Cube has also released monolingual BERT models for other Indic languages including Kannada, Telugu, Malayalam, Tamil, Gujarati, Assamese, Odia, Bengali, and Punjabi. These models aim to bring the power of BERT to a wider range of languages, particularly those that may not have as many resources available for NLP tasks. We will be using the Telugu version of this model.

# Chapter 4

## Methodology

### 4.1 Dataset Description

This dataset is taken from the research paper TeQuAD Rakesh et al. [2]. This dataset mainly contains 82k triples split into 66k train triples and 16k test triples each of format as follows:

- Context in Telugu
- Question in Telugu
- Answer in Telugu

### 4.2 Data Preprocessing

#### 4.2.1 Loading the dataset

For loading the dataset we create a function `get_data()`. This function loads the data from all the text files in UTF-8 format. It combines the context, questions, and answers into a single list of tuples, where each tuple contains a context, a question, and an answer. It splits the combined data into training and validation sets using an 80-20 split. The `train_test_split()` function from the `sklearn` library is used for this purpose. The random state is set to 42 to ensure that the splits are reproducible. It unpacks the split data into separate lists for the training and validation contexts, questions, and answers.

#### 4.2.2 Checking Context and Question Distributions

During tokenization, we need to make sure that we are truncating and padding the tokens at same appropriate length. To find this appropriate length, we draw the length distribution of Context and Questions.

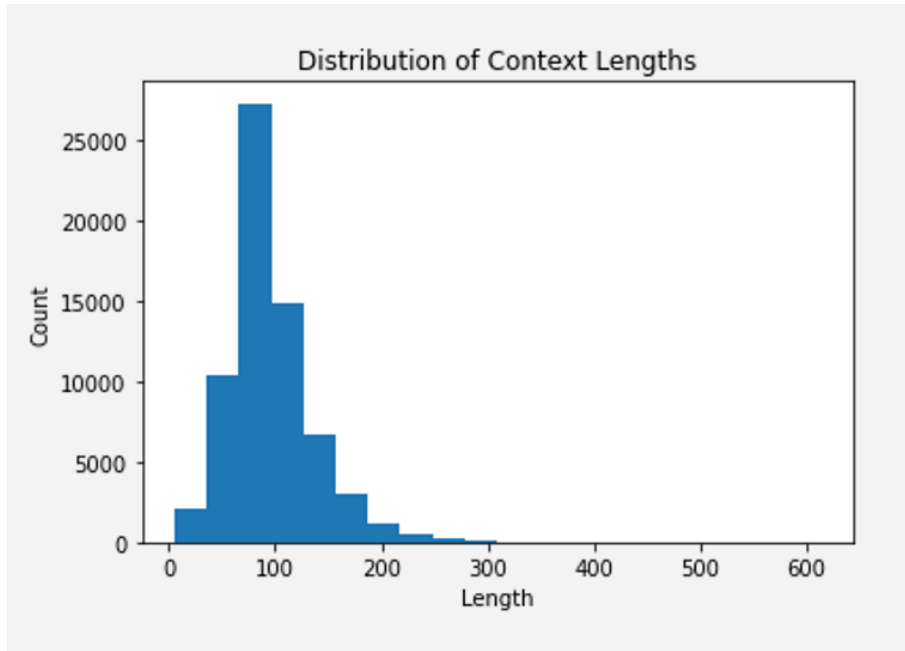


Fig 3.1: Distribution of Context Lengths

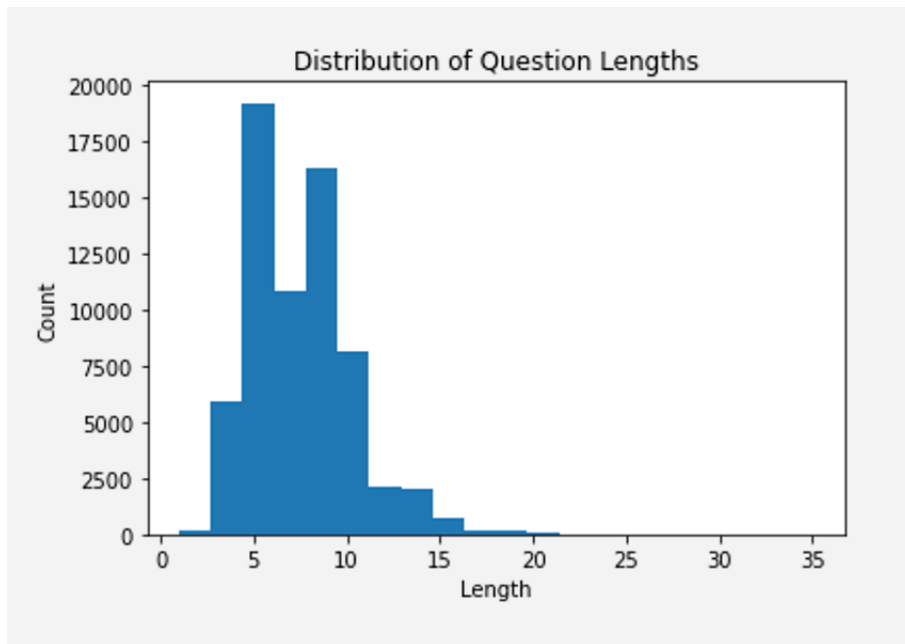


Fig 3.2: Distribution of Question Lengths

As we can see from the distributions, majority of data points have a context length of less than 300 and a question length of less than 20. So we can pick 300 as the `max_length` to truncate.

### 4.2.3 Tokenization

To get the tokenizer, we use the hugging face's AutoTokenizer. A tokenizer is responsible for preprocessing text into an array of numbers as inputs to a model. Multiple rules govern the tokenization process, including how to split a word and at what level words should be split.

A pretrained tokenizer with the AutoTokenizer.from\_pretrained() method. This downloads the vocab a model was pretrained with.

```
from transformers import AutoTokenizer, AutoModelForSeq2SeqLM, AutoModel
tokenizer = AutoTokenizer.from_pretrained("l3cube-pune/telugu-bert")
```

Fig 3.3: Download the tokenizer

Then pass your text to the tokenizer

```
train_encodings = tokenizer(train_questions, train_context, max_length = MAX_LENGTH, truncation=True, padding=True)
valid_encodings = tokenizer(val_questions, val_context, max_length = MAX_LENGTH, truncation=True, padding = True)
```

Fig 3.4: Pass the train\_data and val\_data to the tokenizer

The tokenizer returns a dictionary with three important items:

- input\_ids are the indices corresponding to each token in the sentence.
- attention\_mask indicates whether a token should be attended to or not.
- token\_type\_ids identifies which sequence a token belongs to when there is more than one sequence.

and we can decode the encoded input using tokenizer.decode.

### 4.2.4 Appending Start and End Positions of Answer to Encoded data

Now what we need to send to whatever model we are training apart from question and context is the answer. But we need not send the textual form of the answer. Instead, we can send the span indices of the answer to the model. These span indices mainly consist of the start\_position and end\_position of the answer.

---

**Algorithm 1** Function to return start and end indices of an answer in train\_dataset

---

**Require:** *idx*: Index of the answer

**Ensure:** Start and end indices of the answer

```
1: ret_start  $\leftarrow$  0
2: ret_end  $\leftarrow$  0
3: answer_encoding  $\leftarrow$  Tokenize the answer using the tokenizer with maximum length
   MAX_LENGTH, truncating and padding if necessary
4: for a in range (length of train_encodings['input_ids'] [idx] -
   length of answer_encoding['input_ids']) do
5:   match  $\leftarrow$  True
6:   iter  $\leftarrow$  0
7:   for i in range (1, length of answer_encoding['input_ids] - 1) do
8:     iter  $\leftarrow$  i
9:     if (answer_encoding['input_ids'] [i]  $\neq$  train_encodings['input_ids'] [idx] [a + i])
       then
10:      match  $\leftarrow$  False
11:      break
12:    end if
13:  end for
14:  if match then
15:    ret_start  $\leftarrow$  a + 1
16:    ret_end  $\leftarrow$  a + iter + 1
17:    break
18:  end if
19: end for
20: return (ret_start, ret_end)
```

---

The above algorithm finds the start and the end positions of train\_answers in train\_context. An example of what the output looks like is as follows.

```
test_rec=99

z,x = ret_Answer_start_and_end_train(test_rec)
print(z, x)

predict_answer_tokens = train_encodings.input_ids[test_rec][z : x]
print(tokenizer.decode(predict_answer_tokens))
print(train_answers[test_rec]['text'])
print(tokenizer.decode(train_encodings['input_ids'][test_rec]))
```

24 29  
అలెగ్జాండర్ లాడియోగిన్  
అలెగ్జాండర్ లాడియోగిన్  
[CLS] ఎవరు 1874 లో రష్యాలో ప్రకాశించే కాంతి బల్బు పేటెంట్ చేసారు? [SEP] 1872 లో, రష  
రియు 1874 లో రష్యాన్ పేటెంట్లు పొందారు. అతను ఒక గ్లాస్ రిసీవర్ రెండు కార్బన్ రాడ్ల కొలిపి  
నిండిపోయింది, ఎలక్ట్రిక్ ఛార్జ్ పర్యవేక్షించి, దీని వలన విద్యుత్తు మొదటిసారి వినియోగించిన రెండవ  
పేరును అలెగ్జాండర్ డి లాడిగూయిస్ మార్చుకున్నాడు మరియు క్రోమియం, ఇరిడియం, రోడియం,  
ప్రకాశవంతమైన దీపాలకు దరఖాస్తు మరియు దరఖాస్తు చేసుకున్నారు మరియు ఒక మాల్టిఫ్లెక్స్ ఫిల్  
న్యాయమైనది. [SEP] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PA  
[PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD]  
[PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD]  
[PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD]

Fig 3.5: Example Usage of ret\_Answer\_start\_and\_end\_train

This can similarly be applied to `valid_encodings`. Now we can append start positions and end positions to train and valid encodings.

---

**Algorithm 2** Update `train_encodings` with start and end positions

---

```

1: Initialize empty lists: start_positions, end_positions
2: Initialize counter ctr  $\leftarrow$  0
3: for h in range (len(train_encodings['input_ids'])) do
4:   Compute s, e  $\leftarrow$  ret_Answer_start_and_end_train(h)
5:   Append s to start_positions
6:   Append e to end_positions
7:   if s == 0 then
8:     Increment ctr by 1
9:   end if
10: end for
11: Update train_encodings with start_positions and end_positions
12: Print ctr

```

---

Once the start and end positions are appended to valid and `train_encodings`. The data preprocessing is complete.

The dictionary keys of train or `valid_encodings` are as follows:

- `input_ids`
- `attention_mask`
- `token_type_ids`
- `start_positions`
- `end_positions`

## 4.3 Training the model

First and foremost, we need to import the model. We can do that by hugging face using various forms of `AutoModel`.

```
bert_model = AutoModel.from_pretrained("ai4bharat/indic-bert")
```

Fig 3.6: Importing IndicBERT from hugging-face.

Since all the models that are used now (IndicBERT, RoBERTa, L3cube, mBERT) are all based on BERT. We see that the dimensionality of the encoder and pooler layer is the same figure (768). To be able to use these models for our question-answering task, We need to add additional layers on top of the model so that we get the appropriate output.

Also, we need to change the forward function to fit the model we are running. In order to do this, in implementation we can create a question-answering model class. The pseudo-code for the implementation is as below

Initialization:

1. Initialize the BERT model
2. Initialize the dropout layer with a dropout rate of 0.1
3. Initialize linear layer 1 with input size  $768*2$  and output size  $768*2$
4. Initialize linear layer 2 with input size  $768*2$  and output size 2
5. Combine layers in a sequential stack: dropout, linear layer 1, LeakyReLU activation

Forward Pass:

1. Get model output from BERT with input\_ids, attention\_mask, and token\_type\_ids
2. Extract hidden states from the model output
3. Concatenate last and third to last hidden states
4. Pass the concatenated output through the sequential stack to get logits
5. Split logits into start\_logits and end\_logits
6. Squeeze the last dimension of start\_logits and end\_logits
7. Return start\_logits, end\_logits

The model and the dataset are now set up. Now we can train the model.

The training parameters are

- **Batch-Size:** 16
- **Epochs:** 3
- **Weight Optimiser:** Adam Optimiser
- **Learning rate:**  $2 \times 10^{-5}$
- **Weight decay:**  $2 \times 10^{-2}$
- **Scheduler:** Exponential-LR



# Chapter 5

## Results

### 5.1 XLM-RoBERTa

#### 5.1.1 Results

The model ran for about 3 epochs. At the end of three epochs, the training accuracy was around 59.694 % with a train loss of 1.275 and validation accuracy of 55.304 %.

The test experiments were run on the 20% of the dataset which was split at the start randomly.

We mainly calculated the F1\_score, Exact Match and the Partial Match.

F1 Score	Exact Match	Partial Match
0.5753585240145673	0.38454088735548697	0.736335572907209

Table 5.1: Scores Table

#### 5.1.2 Cases where the model fails

The following examples show why the model has such low EM scores.

##### Case 1: Bad Exact Match due to Underpredicting Partial Match.

**Telugu-Context:** కేంద్రీయ భాష ప్రామాణిక భాషగా భావించబడుతుంది మరియు అత్యధిక సంఖ్యలో మాట్లాడేవారు ఉన్నారు. ఇది బార్సిలోనా ప్రావిన్స్ యొక్క జనసాంద్రత కలిగిన ప్రాంతాలలో, తారగానో ప్రావిన్స్ యొక్క తూర్పు భాగంలో మరియు గిరోనా ప్రావిన్స్ లో చాలా ప్రాంతాలలో మాట్లాడబడుతుంది.

**Translated English Version:** The central language is considered the standard language and has the largest number of speakers. It is spoken in densely populated areas of the province of Barcelona, in the eastern part of the province of Tarragona, and in most areas of the province of Girona.

**Telugu-Question:** ఏ ప్రామాణిక ప్రామాణిక ఉద్భావణను కలిగి ఉంటుంది?

**Translated English Version:** Which Language is the standard language?

**Expected Answer:** కేంద్రీయ భాష (Central Language)

**Actual Answer:** కేంద్రీయ (Central)

## Case 2: Bad Exact Match due to Overpredicting Partial Match.

**Telugu-Context :** పోటీదారులు కనీసం మూడు సెట్ల కోతలు ద్వారా వెళతారు. మొదట ప్రదర్శనకారుల ముందు కొన్ని ఇతర పోటీదారులతో క్లుప్తంగా ఆడిషన్ ఉంది , ఇందులో ప్రదర్శన యొక్క నిర్మాతలలో ఒకరు కూడా ఉండవచ్చు. ఆడిషన్లు ప్రతి నగరంలో 10,000 మించగలిగినప్పటికీ , వాటిలో కొన్ని వందల మాత్రమే ఆరంభ దశలోనే ఆడిషన్లు జరుగుతాయి. విజయవంతమైన పోటీదారులు అప్పుడు నిర్మాతల ముందు పాడతారు , ఇక్కడ ఎక్కువ కట్ చేయవచ్చు. అప్పుడు మాత్రమే వారు న్యాయమూర్తుల ముందు ఆడిషన్లు వెళ్ళవచ్చు , ఇది టెలివిజన్లో చూపించిన ఏకైక ఆడిషన్ స్టేజి. న్యాయమూర్తులు ఎంపిక చేసిన వారు హాలీవుడ్కు పంపబడ్డారు. ప్రతి నగరంలో 10-60 మంది మధ్య హాలీవుడ్కు ఇది ఉపయోగపడుతుంది.

**Translated English Version:** Competitors go through at least three sets of cuts. First the performers have a brief audition with some of the other contestants, which may include one of the show's producers. Although auditions can exceed 10,000 in each city, only a few hundred of them are auditioned at the initial stage. Successful contestants then sing in front of producers, where more cuts can be made. Only then can they go on to audition in front of the judges, the only audition stage shown on television. Those chosen by the judges were sent to Hollywood. It serves Hollywood between 10-60 people in each city.

**Telugu-Question:** హాలీవుడ్కు ముందు ఎంత రౌండ్లు పోటీ పడుతున్నాయి ?

**Translated English Version:** How many sets of cuts do they go through before reaching Hollywood?

**Expected Answer:** మూడు (Three)

**Actual Answer:** కనీసం మూడు సెట్ల కోతలు (Atleast three sets of cuts)

## Case 3: Bad Exact Match due to Completely wrong prediction

**Telugu-Context :** నగరం సద్నాలుగో శతాబ్దం మధ్యలో తీవ్రమైన సమస్యల ద్వారా వెళ్ళింది. ఒక వైపున 1348 యొక్క నల్ల మరణం మరియు తరువాతి సంవత్సరాల్లో అంటురోగాల ద్వారా ప్రజల యొక్క శిథిలమయ్యాయి - మరియు మరొకదానిపై , యుద్ధాల సేరీస్ మరియు తరువాత జరిగిన అల్లర్లు. వీరిలో యూనియన్ యొక్క యుద్ధం , సామ్రాజ్య రాజధానిగా వాలెన్సియా సేత్త్యత్వంలో రాచరికం యొక్క అతిక్రమణకు వ్యతిరేకంగా ఒక పోరుడు తిరుగుబాటు - మరియు 1363 లో కరేలియన్ దాడులను అడ్డుకోవాలనే కొత్త గోడను ఆగ్రహానికి పెంచడంతో , మరియు 1364. ఈ సంవత్సరాల్లో నగర-కైస్తువుని , యూదు మరియు ముస్లింను ఆక్రమించిన మూడు వర్గాల సహజీవనం చాలా వివాదాస్పదంగా ఉంది. వాటర్నంట్ చుట్టూ ఉన్న ప్రాంతాన్ని ఆక్రమించిన యూదులకు ఆర్థికపరంగా మరియు సామాజికంగా అభివృద్ధి చెందింది మరియు వారి త్రైమాసికంలో పోరుగు పారిష్ల వ్యయంతో క్రమంగా దాని సరిహద్దులను విస్తరించింది. అదే సమయంలో , ఈ విజయం తర్వాత నగరంలో మిగిలిపోయిన ముస్లింలు ప్రస్తుత మార్కెట్ మార్కెట్లో నోర్త్ ప్రక్కన అసంపూర్ణ పోరుగు ప్రాంతంలో నివసించారు. 1391 లో ఒక అనియంత్రిత అనారోగ్యం యూదుల త్రైమాసికంపై దాడి చేసింది , దాని వాస్తవిక అదృశ్యం మరియు దాని మనుగడలో ఉన్న సభ్యులను కైస్తువ మతానికి బలవంతంగా మార్చడానికి దారితీసింది. 1456 లో జనాభాలో ఇదే తరహా అల్లర్ల సమయంలో ముస్లిం క్వార్టర్ దాడి జరిగింది , అయితే పరిణామాలు చిన్నవి.

**Translated English Version:** The city went through serious troubles in the middle of the fourteenth century. On the one hand was the Black Death of 1348 and the devastation of the people by epidemics in the following years – and on the other, a series of wars and riots that followed. These included the War of the Union, a citizen revolt against the encroachment of the monarchy led by Valencia as the imperial capital – and the raising of a new wall to ward off Karelian attacks in 1363, and 1364. The coexistence of the three factions that occupied the city—Christian, Jewish, and Muslim—was highly controversial during these years. The Jews who occupied the area around the waterfront prospered economically and socially and gradually expanded its boundaries at the expense of neighboring parishes in their quarter. At the same time, the Muslims who remained in the city after this conquest lived in an incomplete neighborhood next to Sorle in the present market square. In 1391 an uncontrolled illness attacked the Jewish quarter, leading to its virtual disappearance and the forced conversion of its surviving members to Christianity. In 1456, during a similar riot among the population, the Muslim quarter was raided, but the consequences were minor.

**Telugu-Question:** యూదుల విభాగం ఎప్పుడు దాడికి గురైంది ?

**Translated English Version:** When was the Jewish section attacked?

**Expected Answer:** 1391 (1391)

**Actual Answer:** 1456 (1456)

## 5.2 L3-Cube

### 5.2.1 Results

The model ran for about 3 epochs. At the end of three epochs, the training accuracy was around 62.84 % with a train loss of 1.256 and validation accuracy of 54.67 %.

The test experiments were run on the 20% of the dataset which was split at the start randomly.

We mainly calculated the F1\_score, Exact Match and the Partial Match.

F1 Score	Exact Match	Partial Match
0.6998489065492597	0.44561467223533685	0.603111191816476

Table 5.2: Scores Table

### 5.2.2 Cases where the model fails

#### Case 1: Bad Exact Match due to Wrong Prediction.

**Telugu-Context:** నగరం యొక్క పరిపాలనా సంస్థల పరిధులలో, పరిమాణ క్రమంలో : హైదరాబాద్ పోలీస్ ప్రాంతం , హైదరాబాద్ జిల్లా , జిమ్క ప్రాంతం ( " హైదరాబాద్ నగరం " ) మరియు హైదరాబాదు మెట్రోపాలిటన్ డెవలప్మెంట్ అథారిటీ ( hmda ) కింద ఉన్న ప్రాంతం. hmda అనేది ఒక అపోలాపియల్ పట్టణ ప్రణాళికా యంత్రాంగం , ఇది ghmc మరియు దాని శివారు ప్రాంతాలను కలిగి ఉంది , నగరంలో చుట్టుముట్టబడిన ఐదు జిల్లాలలో 54 మండలాలకు విస్తరించింది. ఇది ghmc మరియు సబర్బన్ మున్సిపాలిటీల అభివృద్ధి కార్యకలాపాలను సమన్వయపరుస్తుంది మరియు హైదరాబాద్ మెట్రోపాలిటన్ నీటి సరఫరా మరియు మురుగునీటి బోర్డు ( hmwssb ) వంటి సంస్థల పరిపాలనను నిర్వహిస్తుంది.

**Translated English Version:** The administrative bodies of the city include, in order of size: Hyderabad Police area, Hyderabad district, JIMK area ("Hyderabad City") and the area under the Hyderabad Metropolitan Development Authority (hmda). hmda is an apocalyptic town planning agency that includes ghmc and its suburbs, spread over 54 mandals in five districts surrounding the city. It coordinates the development activities of the ghmc and suburban municipalities and manages the administration of institutions such as the Hyderabad Metropolitan Water Supply and Sewerage Board (hmwssb).

**Telugu-Question:** ప్రధానంగా బాధ్యతాయుతంగా hmda ఏమిటి ?

**Translated English Version:** What is hmda primarily responsible for?

**Expected Answer:** పట్టణ ప్రణాళికా (Urban planning)

**Actual Answer:** హైదరాబాదు మెట్రోపాలిటన్ డెవలప్మెంట్ అథారిటీ (Hyderabad Metropolitan Development Authority)

## Case 2: Bad Exact Match due to Underpredicting Partial Match.

**Telugu-Context:** ఈ విషయంలో, అష్కెజీ యొక్క కొంటర్ సెఫర్డిక్, ఎందుకంటే చాలా మంది అష్కెజీ సంప్రదాయ యూదులు సెఫర్డిక్ రాబినికల్ అధికారులను అనుసరిస్తారు, వారు జాతిపరంగా సెఫర్డిక్ లోనే ఉంటారు. సంప్రదాయం ద్వారా, ఒక సెఫర్డిక్ లేదా మిజ్రాహి మహిళ ఒక సనాతన లేదా హర్దిరి అష్కెజీ యూదు కుటుంబంలో పెళ్లి చేసుకుంటాడు, ఆమె పిల్లలు అష్కెజీ యూదులను పెంచుతాడు; ఒక సెఫర్డి లేదా మిజ్రాహి మనిషిని వివాహం చేసుకునే అసహనజీ మహిళ, సెఫర్డిక్ ఆచరణలో తీసుకోవాలని భావిస్తున్నారు మరియు పిల్లలు సెఫర్డిక్ గుర్తింపును వారసత్వంగా పొందుతారు, అయితే అనేక కుటుంబాలు రాజీ పడతాయి. ఒక మార్పిడి సాధారణంగా అతని లేదా ఆమెని మార్పిడి బాత్ దిన్ యొక్క అభ్యాసాన్ని అనుసరిస్తుంది. ఇస్రాయేల్, ఉత్తర అమెరికా, మరియు ఇతర ప్రదేశాలలో ప్రపంచవ్యాప్తంగా యూదులను ఏకీకరణ చేయటంతో, ఒక అసహనజీ యూదు మతపరమైన నిర్వచనం ప్రత్యేకంగా సాంప్రదాయక జడాయిజం వెలుపల అస్పష్టంగా ఉంది.

**Translated English Version:** In this respect, Ashkekezi's counterpart is Sephardic, because most Ashkekezi traditional Jews follow Sephardic rabbinical authorities, whether they are ethnically Sephardic or not. By tradition, a Sephardic or Mizrahi woman marries into an orthodox or Hardiri Ashkekazi Jewish family, raising her children Ashkekazi Jews; An Asahanaji woman who marries a Sephardi or Mizrahi man is expected to adopt Sephardic practices and the children inherit Sephardic identity, but many families compromise. A conversion usually follows the practice of his or her converted bath din. With the integration of Jews around the world in Israel, North America, and elsewhere, the religious definition of an Intolerant Jew is particularly vague outside of orthodox Judaism.

**Telugu-Question:** ఒక మార్పిడి సాధారణంగా ఏ అభ్యాసాన్ని అనుసరిస్తుంది?

**Translated English Version:** What practice does a conversion generally follow?

**Expected Answer:** అతని లేదా ఆమెని మార్పిడి బాత్ దిన్ యొక్క అభ్యాసాన్ని (The practice of bath din that changed him or her)

**Actual Answer:** బాత్ దిన్ (Bath din)

## 5.3 Indic-BERT

### 5.3.1 Results

The model ran for about 5 epochs. At the end of three epochs, the training accuracy was around 47.21 % with a train loss of 1.975 and validation accuracy of 27.6 %.

The test experiments were run on the 20% of the dataset which was split at the start randomly.

We mainly calculated the F1\_score, Exact Match and the Partial Match.

F1 Score	Exact Match	Partial Match
0.44365046866222424	0.26100720295381635	0.35064463410205193

Table 5.3: Scores Table

### 5.3.2 Cases where the model fails

#### Case 1: Bad Exact Match due to Wrong Prediction.

**Telugu-Context :** సమ స్మిత్ మరియు రెగ్యులర్ సహకారి జిమ్మీ నాప్స్ చిత్రం యొక్క శీర్షిక సేవధ్యం " స్మైల్ ఆన్ ది వాల్ " ప్రాసెస్ అప్ డి 2015 లో ప్రకటించబడింది , ఈ సినీమా కోసం స్మిత్ దీనిని ప్రదర్శించారు. స్మిత్ పాట ఒక సెషన్ లో కలిసి వచ్చింది మరియు అతను మరియు napes ఒక డెమో రికార్డింగ్ ముందు అరగంటలో అది రాశాడు. నాణ్యత సంతృప్తి , డెమో తుది విడుదలలో ఉపయోగించారు.

**Translated English Version:** It was announced in September 2015 that Sam Smith and regular collaborator Jimmy Knaps wrote the film's title theme, "Smile on the Wall", which Smith performed for the film. Smith came up with the song in one session and wrote it in half an hour before he and Napes recorded a demo. Quality satisfaction, demo used in final release.

**Telugu-Question:** ఈ థీమ్ యొక్క వాస్తవ చిత్రం వాస్తవ చిత్రంలో ఉపయోగించబడింది ?

**Translated English Version:** What version of this theme used in the film?

**Expected Answer:** డెమో (Demo)

**Actual Answer:** సమ స్మిత్ (Sam Smith)

**Telugu-Context :** ఇతర స్థానిక కళాశాలలు మరియు విశ్వవిద్యాలయాలు కన్సోర్షియా యూనివర్సిటీ వార్ ఆర్బర్, లూథరన్ లిబరల్ ఆర్ట్స్ సంస్థ; పీనిక్స్ విశ్వవిద్యాలయం యొక్క క్యాంపస్; మరియు క్లియర్ విశ్వవిద్యాలయం, ఒక ప్రైవేట్ వ్యాపార పాఠశాల. ఉమ్మడి కమ్యూనిటీ కళాశాల పొరుగున ఉన్న అర్బర్ టౌన్షిప్లో ఉంది. 2000 లో, అవాన్ మారియా పాఠశాల చట్టం, డొమినోస్ పిజ్జా వ్యవస్థాపకుడు టామ్ ముగ్గాన్ స్థాపించిన రోమన్ కాథలిక్ లా స్కూల్, ఈశాన్య దిన ఆర్నాల్డ్ ప్రారంభించబడింది, కాని పాఠశాల 2009 లో అవే మారియా, ఫ్లోరిడా మరియు థామస్ m కు తరలించబడింది. కోయలి లా స్కూల్ పాఠశాలలో ఒక ప్రాంగణ ప్రాంగణం కోసం పూర్వపు మరీ భవనాలను కొనుగోలు చేసింది.

**Translated English Version:** Other local colleges and universities include Concordia University in War Arbor, a Lutheran liberal arts institution; campus of the University of Phoenix; and Clear University, a private business school. Saltwater Community College is located in neighboring Arbor Township. In 2000, Avon Maria School of Law, a Roman Catholic law school founded by Domino's Pizza founder Tom Maughan, opened in Northeast Dinaarch, but the school moved in 2009 to Avon Maria, Florida, and Thomas m. Koyali Law School acquired the former Mari buildings for a campus in the school.

**Telugu-Question:** ఒక రోమన్ కాథలిక్ పాఠశాల అక్కడ నుండి వెళ్ళిన తరువాత పాఠశాల ఏవి మారియా భవనాన్ని కొనుగోలు చేసింది?

**Translated English Version:** Which school bought the Avi Maria building after a Roman Catholic school moved out?

**Expected Answer:** కోయలి లా స్కూల్ (Koyali Law School)

**Actual Answer:** డొమినోస్ పిజ్జా వ్యవస్థాపకుడు టామ్ ముగ్గాన్ స్థాపించిన రోమన్ కాథలిక్ లా స్కూల్, ఈశాన్య దిన ఆర్నాల్డ్ ప్రారంభించబడింది, కాని పాఠశాల 2009 లో అవే మారియా, ఫ్లోరిడా మరియు థామస్ m

(The Roman Catholic law school, founded by Domino's Pizza founder Tom Mogghan, opened in Northeast Dinaarch, but the school moved to Ave Maria, Florida, and Thomas m in 2009.)

## Case 2: Bad Exact Match due to Partial Match.

**Telugu-Context :** మార్చి 2005 లో , భద్రతా మండలి అధికారికంగా దార్ఫూర్లో అంతర్జాతీయ క్రిమినల్ కోర్టు యొక్క ప్రాసిక్యూటర్లు నివేదించింది , కమిషన్ రిపోర్టును పరిగణనలోకి తీసుకున్నప్పటికీ , ఏ నిర్దిష్ట నేరాలను పేర్కొనకుండా. సెక్యూరిటీ కౌన్సిల్ యొక్క రెండు శాశ్వత సభ్యులు , యునైటెడ్ స్టేట్స్ మరియు చైనా , రిఫరల్ రిజల్యూషన్ ఓటు నుండి తప్పుకున్నారు. సెక్యూరిటీ కౌన్సిల్ తన నాల్గవ నివేదిక ప్రకారం , ప్రాసిక్యూటర్ గుర్తించిన వ్యక్తులు " [ భద్రతా మండలి తీర్మానం 1593 లో ] మానవత్వం మరియు యుద్ధ నేరాలకు వ్యతిరేకంగా నేరాలకు పాల్పడినట్లు విశ్వసించడానికి సహేతుక మైదానాలు " కనుగొన్నారు , కానీ విచారణకు తగినంత సాక్ష్యం కనుగొనలేదు జెనోసైడ్ కోసం.

**Translated English Version:** In March 2005, the Security Council formally reported to the Prosecutor of the International Criminal Court in Durbar, although the commission considered the report, without specifying any specific crimes. Two permanent members of the Security Council, the United States and China, abstained from voting on the referral resolution. According to the Security Council's fourth report, the prosecutor found "reasonable grounds to believe that the persons identified [in Security Council resolution 1593] committed crimes against humanity and war crimes", but did not find sufficient evidence to prosecute for genocide.

**Telugu-Question:** భద్రతా మండలి అధికారికంగా దార్ఫూర్లో పరిస్థితిని ఎవరికి అప్పగించింది ?

**Translated English Version:** To whom has the Security Council formally entrusted the situation in Darfur?

**Expected Answer:** అంతర్జాతీయ క్రిమినల్ కోర్టు యొక్క ప్రాసిక్యూటర్లు (Prosecutor of the International Criminal Court)

**Actual Answer:** ప్రాసిక్యూటర్లు (Prosecutor)

## 5.4 m-BERT (Prior-Work)

### 5.4.1 Results

This model was actually trained in last semester. But it was tested against a very small test dataset (only 1000 records) and gave an accuracy of 65.9% FM and 39% EM. So this time we run it against our test dataset with around 16k records. The model ran for about 3 epochs. At the end of three epochs, the training accuracy was around 58.31 % with a train loss of 1.40 and a validation accuracy of 50.9 %.

The test experiments were run on the 20% of the dataset which was split at the start randomly.

We mainly calculated the F1\_score, Exact Match and the Partial Match.

F1 Score	Exact Match	Partial Match
0.5936724595866701	0.35597118818473455	0.6048060044791478

Table 5.4: Scores Table

### 5.4.2 Cases where the model fails

#### Case 1: Bad Exact Match due to Wrong Prediction.

**Telugu-Context :** 1960 వ దశకంలో ఇరానియన్ సినిమాకి ఒక ముఖ్యమైన దశాబ్దం ఉంది , 60 ల ప్రారంభంలో సగటున సంవత్సరానికి 25 వాణిజ్య చిత్రాలు నిర్మించబడ్డాయి , ఇది దశాబ్దం చివరి నాటికి 65 కి పెరిగింది. మెలోడ్రామా మరియు థ్రిల్లర్లపై దృష్టి సారించిన ఉత్పత్తిలో ఎక్కువ భాగం. 1969 లో మాసౌద్ కిమియా మరియు డారిష్ మేహర్జియి దర్శకత్వం వహించిన చిత్రాల కైజర్ మరియు ఆవు చిత్రాల ప్రదర్శనతో , ప్రత్యామ్నాయ చలనచిత్రాలు చలన చిత్ర పరిశ్రమలో తమ చోదాను స్థాపించాయి. 1954 లో ప్రారంభమైన ఒక ఫిల్మ్ ఫెస్టివల్ గోల్రిజన్ ఫెస్టివల్ పరిధిలో నిర్వహించడానికి ప్రయత్నాలు జరిగాయి , 1969 లో సెపాస్ ఫెస్టివల్ రూపంలో పండ్లు కోట్టాయి. ఈ ప్రయత్నాలు 1973 లో టెహ్రన్ వరల్డ్ ఫెస్టివల్ ఏర్పడడానికి కారణమయ్యాయి.

**Translated English Version:** The 1960s was an important decade for Iranian cinema, with an average of 25 commercial films produced annually in the early 60s, rising to 65 by the end of the decade. Much of the production focused on melodrama and thrillers. Alternative films established their status in the film industry in 1969 with the release of the films Kaiser and Cow directed by Masoud Kimia and Darish Mehrijui. Beginning in 1954, efforts were made to organize a film festival under the Golrizon Festival, which bore fruit in 1969 in the form of the Sepas Festival. These efforts led to the creation of the Tehran World Festival in 1973.

**Telugu-Question:** ఇరాన్లో 1960 ల చివరినాటికి ఎన్ని వాణిజ్య సినిమాలు సంవత్సరానికి సగటున ఉత్పత్తి చేయబడ్డాయి ?

**Translated English Version:** How many commercial films were produced per year on average in late 1960s in Iran?

**Expected Answer:** 65

**Actual Answer:** 25



## Case 2 : Bad Exact Match due to Overpredicting Partial Match.

**Telugu-Context:** సాంప్రదాయకంగా, సాధారణ కార్బోహైడ్రేట్లు త్వరితంగా శోషించబడుతున్నాయని భావిస్తున్నారు, అందువల్ల సంక్లిష్ట కార్బోహైడ్రేట్లు కంటే రక్తం-గ్లూకోజ్ స్థాయిలు వేగంగా పెరుగుతాయి. అయితే ఇది ఖచ్చితమైనది కాదు. కొన్ని సాధారణ కార్బోహైడ్రేట్లు (ఉదా. ఫ్రక్టోజ్) వివిధ జీవక్రియ మార్గాలు (ఉదా., ఫ్రక్టోలిసిస్) ను అనుసరిస్తాయి, దీని వలన గ్లూకోజ్కు ఒక పాక్షిక క్యాటాబోలిజం మాత్రమే వస్తుంది, అయితే, సంక్లిష్టంగా, అనేక సంక్లిష్ట కార్బోహైడ్రేట్లు సాధారణ కార్బోహైడ్రేట్లకు సమాన స్థాయిలో జీర్ణమవుతాయి. గ్లూకోజ్ రక్తప్రవాహంలోకి ప్రవేశించడం ద్వారా ఇన్సులిన్ ఉత్పత్తిని ఉత్తేజితం చేస్తుంది, ఇది ప్యాంక్రియాస్లో బీటా కణాల ద్వారా రహించబడింది.

**Translated English Version:** Traditionally, simple carbohydrates have been thought to be absorbed more quickly and therefore raise blood-glucose levels faster than complex carbohydrates. But this is not accurate. Some simple carbohydrates (e.g. fructose) follow different metabolic pathways (e.g., fructolysis) resulting in only a partial catabolism to glucose, but, in contrast, many complex carbohydrates are digested at the same rate as simple carbohydrates. Glucose entering the bloodstream stimulates the production of insulin, which is absorbed by beta cells in the pancreas.

**Telugu-Question:** సాంప్రదాయకంగా వేగంగా రక్తనాళాల గ్లూకోజ్ స్థాయిలు పెరగడంతో శోషించబడిందని నమ్మకం ఏమిటి?

**Translated English Version:** What is traditionally believed to be rapidly absorbed as blood glucose levels rise?

**Expected Answer:** సాధారణ కార్బోహైడ్రేట్లు (simple carbohydrates)

**Actual Answer:** సాధారణ కార్బోహైడ్రేట్లు త్వరితంగా శోషించబడుతున్నాయని భావిస్తున్నారు, అందువల్ల సంక్లిష్ట కార్బోహైడ్రేట్లు కంటే రక్తం - గ్లూకోజ్ స్థాయిలు వేగంగా పెరుగుతాయి. అయితే ఇది ఖచ్చితమైనది కాదు

(Simple carbohydrates are thought to be absorbed more quickly and therefore raise blood-glucose levels faster than complex carbohydrates. But this is not accurate)

**Telugu-Context:** రూపకల్పన ద్వారా, దాని లోపలికి సరిగ్గా ఒక USB ప్లగ్ని ఇన్సర్ట్ చేయడం కష్టం. USB స్పెసిఫికేషన్ ప్రకారం, USB ప్లగ్ "topside" లో వుండాలి, ఇది "... సులభంగా వినియోగదారు గుర్తింపును అందిస్తుంది మరియు సమయంలో అమరికను సులభతరం చేస్తుంది." వివరణ కూడా "సిఫార్సు" తయారీదారు యొక్క లోగో "(డ్రైలాగ్రాఫ్) చెక్కినది "కానీ టెక్స్ట్ పేర్కొనబడలేదు) USB ఐకాన్ యొక్క ఎదురుగా ఉంది. ఈ వివరణ ప్రతి ఐటెమ్కు ప్రక్కనే ఉన్నది, ఈ సందర్భంలో ఐకాన్ ఐకాన్లో కనిపించేలా అనుమతించేలా చేయాలి. "అయితే ఈ వివరణ, పరికరం యొక్క ఎత్తును పరిగణించదు యూజర్ యొక్క కన్ను స్థాయి ఎత్తుకు, కాబట్టి ఒక డెస్క్ మీద కంప్యూటర్కు జతచేయబడినప్పుడు "కనిపించే" కేబుల్ యొక్క వైపు వినియోగదారు నిలబడినా లేదా మోకరిస్తేనా అనే దానిపై ఆధారపడి ఉంటుంది.

**Translated English Version:** By design, it is difficult to properly insert a USB plug into it. According to the USB specification, the USB plug should be on the "topside", which "...provides easy user identification and facilitates alignment during mating." The description also "recommends" the manufacturer's logo ("engraved" in the diagram but not mentioned in the text) on the opposite side of the USB icon. This description should be located next to each item, allowing the icon to be visible in the bag. "However, this description does not consider the height of the device of the user. At eye level height, so when attached to a computer on a desk the "visible" side of the cable depends on whether the user is standing or kneeling.

**Telugu-Question:** ఒక USB ప్లగ్ తో ఏమి కష్టం?

**Translated English Version:** What is difficult with a USB plug?

**Expected Answer:** దాని లోపలికి సరిగ్గా ఒక USB ప్లగ్ని ఇన్సర్ట్ (Insert a USB plug into it properly)

**Actual Answer:** దాని లోపలికి సరిగ్గా ఒక USB ప్లగ్ని ఇన్సర్ట్ చేయడం (Inserting a USB plug right into it)

## 5.5 Cases where RoBERTa and L3-Cube give different results

### 5.5.1 Case 1: Where RoBERTa fails and L3-Cube gives correct answer

#### For RoBERTa

**Telugu-Context:** నగరం సద్నాలుగే శతాబ్దం మధ్యలో తీవ్రమైన సమస్యల ద్వారా వెళ్ళింది. ఒక వైపున 1348 యొక్క నల్ల మరణం మరియు తరువాతి సంవత్సరాల్లో అంటురోగాల ద్వారా ప్రజల యొక్క శిథిలమయ్యాయి - మరియు మరొకదానిపై , యుద్ధాల సేరీస్ మరియు తరువాత జరిగిన అల్లర్లు. వీరిలో యూనియన్ యొక్క యుద్ధం , సామ్రాజ్య రాజధానిగా వాలెన్సియా నేతృత్వంలో రాచరికం యొక్క అతిక్రమణకు వ్యతిరేకంగా ఒక పౌరుడు తిరుగుబాటు - మరియు 1363 లో కరేలియన్ దాడులను అడ్డుకోవాలనే కొత్త గోడను ఆగ్రహానికి పెంచడంతో , మరియు 1364. ఈ సంవత్సరాల్లో నగర-కైస్తవుని , యూదు మరియు ముస్లింను ఆక్రమించిన మూడు వర్గాల సహజీవనం చాలా వివాదాస్పదంగా ఉంది. వాటర్నంట్ చుట్టూ ఉన్న ప్రాంతాన్ని ఆక్రమించిన యూదులకు ఆర్థికపరంగా మరియు సామాజికంగా అభివృద్ధి చెందింది మరియు వారి తైమాసికంలో పొరుగు పారిష్ల వ్యయంతో క్రమంగా దాని సరిహద్దులను విస్తరించింది. అదే సమయంలో , ఈ విజయం తర్వాత నగరంలో మిగిలిపోయిన ముస్లింలు ప్రస్తుత మార్కెట్ మార్కెట్లో నోర్త్ ప్రక్కన అసంపూర్ణ పొరుగు ప్రాంతంలో నివసించారు. 1391 లో ఒక అనియంత్రిత అనారోగ్యం యూదుల తైమాసికంపై దాడి చేసింది , దాని వాస్తవిక అదృశ్యం మరియు దాని మనుగడలో ఉన్న సభ్యులను క్రైస్తవ మతానికి బలవంతంగా మార్చడానికి దారితీసింది. 1456 లో జనాభాలో ఇదే తరహా అల్లర్ల సమయంలో ముస్లిం క్వార్టర్ దాడి జరిగింది , అయితే పరిణామాలు చిన్నవి.

**Translated English Version:** The city went through serious troubles in the middle of the fourteenth century. On the one hand was the Black Death of 1348 and the devastation of the people by epidemics in the following years – and on the other, a series of wars and riots that followed. These included the War of the Union, a citizen revolt against the encroachment of the monarchy led by Valencia as the imperial capital – and the raising of a new wall to ward off Karelian attacks in 1363, and 1364. The coexistence of the three factions that occupied the city—Christian, Jewish, and Muslim—was highly controversial during these years. The Jews who occupied the area around the waterfront prospered economically and socially and gradually expanded its boundaries at the expense of neighboring parishes in their quarter. At the same time, the Muslims who remained in the city after this conquest lived in an incomplete neighborhood next to Sorle in the present market square. In 1391 an uncontrolled illness attacked the Jewish quarter, leading to its virtual disappearance and the forced conversion of its surviving members to Christianity. In 1456, during a similar riot among the population, the Muslim quarter was raided, but the consequences were minor.

**Telugu-Question:** యూదుల విభాగం ఎప్పుడు దాడికి గురైంది ?

**Translated English Version:** When was the Jewish section attacked?

**Expected Answer:** 1391 (1391)|

**Actual Answer:** 1456 (1456)



## For L3-Cube

**Telugu-Context :** నగరం పద్నాలుగు శతాబ్దం మధ్యలో తీవ్రమైన సమస్యల ద్వారా వెళ్ళింది. ఒక వైపున 1348 యొక్క నల్ల మరణం మరియు తరువాతి సంవత్సరాల్లో అంటురోగాల ద్వారా ప్రజల యొక్క శిథిలమయ్యాయి - మరియు మరొకదానిపై , యుద్ధాల సెరీస్ మరియు తరువాత జరిగిన అల్లర్లు. వీరిలో యూనియన్ యొక్క యుద్ధం , సామ్రాజ్య రాజధానిగా వాలెన్సియా నేతృత్వంలో రాచరికం యొక్క అతిక్రమణకు వ్యతిరేకంగా ఒక పౌరుడు తిరుగుబాటు - మరియు 1363 లో కరేలియన్ దాడులను అడ్డుకోవాలనే కొత్త గోడను ఆగ్రహానికి పెంచడంతో , మరియు 1364. ఈ సంవత్సరాల్లో నగర-కైస్తువుని , యూదు మరియు ముస్లింను ఆక్రమించిన మూడు వర్గాల సహజీవనం చాలా వివాదాస్పదంగా ఉంది. వాటర్నంట్ చుట్టూ ఉన్న ప్రాంతాన్ని ఆక్రమించిన యూదులకు ఆర్థికపరంగా మరియు సామాజికంగా అభివృద్ధి చెందింది మరియు వారి తైమాసికంలో పొరుగు పారిష్ల వ్యయంతో క్రమంగా దాని సరిహద్దులను విస్తరించింది. అదే సమయంలో , ఈ విజయం తర్వాత నగరంలో మిగిలిపోయిన ముస్లింలు ప్రస్తుత మార్కెట్ మార్కెట్లో సోదై ప్రక్కన అసంపూర్ణ పొరుగు ప్రాంతంలో నివసించారు. 1391 లో ఒక అనియంత్రిత అనారోగ్యం యూదుల తైమాసికంపై దాడి చేసింది , దాని వాస్తవిక అదృశ్యం మరియు దాని మనుగడలో ఉన్న సభ్యులను కైస్తువ మతానికి బలవంతంగా మార్చడానికి దారితీసింది. 1456 లో జనాభాలో ఇదే తరహా అల్లర్ల సమయంలో ముస్లిం క్వార్టర్ దాడి జరిగింది , అయితే పరిణామాలు చిన్నవి.

**Translated English Version:** The city went through serious troubles in the middle of the fourteenth century. On the one hand was the Black Death of 1348 and the devastation of the people by epidemics in the following years – and on the other, a series of wars and riots that followed. These included the War of the Union, a citizen revolt against the encroachment of the monarchy led by Valencia as the imperial capital – and the raising of a new wall to ward off Karelian attacks in 1363, and 1364. The coexistence of the three factions that occupied the city—Christian, Jewish, and Muslim—was highly controversial during these years. The Jews who occupied the area around the waterfront prospered economically and socially and gradually expanded its boundaries at the expense of neighboring parishes in their quarter. At the same time, the Muslims who remained in the city after this conquest lived in an incomplete neighborhood next to Sorle in the present market square. In 1391 an uncontrolled illness attacked the Jewish quarter, leading to its virtual disappearance and the forced conversion of its surviving members to Christianity. In 1456, during a similar riot among the population, the Muslim quarter was raided, but the consequences were minor.

**Telugu-Question:** యూదుల విభాగం ఎప్పుడు దాడికి గురైంది ?

**Translated English Version:** When was the Jewish section attacked?

**Expected Answer:** 1391 (1391)

**Actual Answer:** 1391 (1391)

## 5.5.2 Case 2: Where L3-Cube fails and RoBERTa gives correct answer

### For RoBERTa

**Telugu-Context:** 1933 లో, గ్రీస్ యొక్క సాంప్రదాయ చర్చి అధికారికంగా ప్రకటించింది ఒక ప్రేమాసన్ ఉండటం మతభ్రష్టత్వము మరియు అందువలన, అతను ప్రశంసలు వరకు, freemasonry పాల్గొన్న వ్యక్తి eucharist పాలుపంచుకోలేరు. ఇది తూర్పు సాంప్రదాయ చర్చి మొత్తం అంతటా సాధారణంగా ధృవీకరించబడింది. ప్రేమాసన్ యొక్క సాంప్రదాయక విమర్శలు రోమన్ కాథలిక్ మరియు ప్రొటెస్టంట్ సంస్కరణలు రెండింటినీ అంగీకరిస్తాయి: " ప్రేమాసన్ అనేది ఒక రహస్యం, రహస్యం మరియు రహస్యం మరియు రహస్యం మరియు హేతువాద సిద్ధాంతం వంటి బోధన మరియు క్రైస్తవ ధోరణికి అనుకూలమైనది కాదు. "

**Translated English Version:** In 1933, the Traditional Church of Greece officially announced that being a Freemason is apostasy and therefore, until he prays, a person who participated in freemasonry cannot participate in the eucharist. This is generally affirmed throughout the entire Eastern Orthodox Church. Traditional criticisms of Freemasonry agree with both the Roman Catholic and Protestant versions: "Freemasonry is a mystery, a mystery and mystery, and a teaching and a Christian tendency, such as mystery and rationalism."

**Telugu-Question:** ఇతర ప్రధాన మతాలు ఒకే నమ్మకాలను కలిగి ఉన్నాయి, ప్రేమాసన్లలో, గ్రీస్ యొక్క సాంప్రదాయ చర్చిగా ?

**Translated English Version:** what are the other major religions which share the same beliefs, in Freemasonry, as the traditional Church of Greece?

**Expected Answer:** రోమన్ కాథలిక్ మరియు ప్రొటెస్టంట్ (Roman Catholic and Protestant)

**Actual Answer:** రోమన్ కాథలిక్ మరియు ప్రొటెస్టంట్ (Roman Catholic and Protestant)

### For L3-Cube

**Telugu-Context:** 1933 లో, గ్రీస్ యొక్క సాంప్రదాయ చర్చి అధికారికంగా ప్రకటించింది ఒక ప్రేమాసన్ ఉండటం మతభ్రష్టత్వము మరియు అందువలన, అతను ప్రశంసలు వరకు, freemasonry పాల్గొన్న వ్యక్తి eucharist పాలుపంచుకోలేరు. ఇది తూర్పు సాంప్రదాయ చర్చి మొత్తం అంతటా సాధారణంగా ధృవీకరించబడింది. ప్రేమాసన్ యొక్క సాంప్రదాయక విమర్శలు రోమన్ కాథలిక్ మరియు ప్రొటెస్టంట్ సంస్కరణలు రెండింటినీ అంగీకరిస్తాయి: " ప్రేమాసన్ అనేది ఒక రహస్యం, రహస్యం మరియు రహస్యం మరియు రహస్యం మరియు హేతువాద సిద్ధాంతం వంటి బోధన మరియు క్రైస్తవ ధోరణికి అనుకూలమైనది కాదు. "

**Translated English Version:** In 1933, the Traditional Church of Greece officially announced that being a Freemason is apostasy and therefore, until he prays, a person who participated in freemasonry cannot participate in the eucharist. This is generally affirmed throughout the entire Eastern Orthodox Church. Traditional criticisms of Freemasonry agree with both the Roman Catholic and Protestant versions: "Freemasonry is a mystery, a mystery and mystery, and a teaching and a Christian tendency, such as mystery and rationalism."

**Telugu-Question:** ఇతర ప్రధాన మతాలు ఒకే నమ్మకాలను కలిగి ఉన్నాయి, ప్రేమాసన్లలో, గ్రీస్ యొక్క సాంప్రదాయ చర్చిగా ?

**Translated English Version:** what are the other major religions which share the same beliefs, in Freemasonry, as the traditional Church of Greece?

**Expected Answer:** రోమన్ కాథలిక్ మరియు ప్రొటెస్టంట్ (Roman Catholic and Protestant)

**Actual Answer:** ( )

## 5.6 Results Summarised

Model	F1 Score	Exact Match	Partial Match
XLM-RoBERTa	0.5753585240145673	0.38454088735548697	0.736335572907209
L3-Cube	0.6998489065492597	0.44561467223533685	0.603111191816476
Indic-BERT	0.44365046866222424	0.26100720295381635	0.35064463410205193
m-BERT	0.5936724595866701	0.35597118818473455	0.6048060044791478

Table 5.5: Combined Scores Table

- **L3-Cube** outperforms all other models with an F1 score of approximately 70%, an exact match score of around 45%, and a partial match score of about 60%. This suggests that L3-Cube is the most effective model among the ones tested for this specific task.
- **XLM-RoBERTa** and **m-BERT** show similar performance, with F1 scores of approximately 58% and 59% respectively. Their exact match scores are around 38% and 36% respectively, and their partial match scores are approximately 74% and 60% respectively. These models provide a good balance between precision and recall but are not as effective as L3-Cube.
- **Indic BERT** has the lowest performance among the models tested, with an F1 score of approximately 44%, an exact match score of around 26%, and a partial match score of about 35%.

## 5.7 Conclusion

In conclusion, the study demonstrates the potential of enhancing Machine Reading Comprehension (MRC) for low-resource languages like Telugu. Among the evaluated models, L3-Cube outperforms the others, including mBERT, XLM-RoBERTa, and Indic-BERT, in terms of F1 score and exact match accuracy. However, XLM-RoBERTa leads to partial match accuracy. This suggests that while L3-Cube is more precise, XLM-RoBERTa might be more robust to partial matches. The results underline the importance of continued exploration and fine-tuning of models for low-resource languages.

# Bibliography

- [1] Priyanka Ravva, Ashok Urlana, Manish Shrivasthava. 2020. AVADHAN: System for Open-Domain Telugu Question Answering. International Institute Of Information Technology, Hyderabad. LTRC lab.
- [2] Rakesh Vemula, Mani Nuthi, and Manish Srivastava. 2022. TeQuAD: Telugu Question Answering Dataset. In Proceedings of the 19th International Conference on Natural Language Processing (ICON), pages 300–307, New Delhi, India. Association for Computational Linguistics.
- [3] Kakwani, Divyanshu & Kunchukuttan, Anoop & Golla, Satish & N.C., Gokul & Bhattacharyya, Avik & Khapra, Mitesh & Kumar, Pratyush. (2020). IndicNLP-Suite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. 4948-4961. 10.18653/v1/2020.findings-emnlp.445.
- [4] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8440–8451, Online. Association for Computational Linguistics.
- [5] Joshi, Raviraj. (2022). L3Cube-HindBERT and DevBERT: Pre-Trained BERT Transformer models for Devanagari based Hindi and Marathi Languages.
- [6] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250.
- [7] Danqi Chen, Jason Bolton, and Christopher D Manning. 2016. A thorough examination of the cnn/daily mail reading comprehension task. arXiv preprint arXiv:1606.02858.
- [8] Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for

- information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.
- [9] Tsung-Yuan Hsu, Chi-Liang Liu, and Hung-yi Lee. 2019. Zero-shot reading comprehension by cross-lingual transfer learning with multi-lingual language representation model. *arXiv preprint arXiv:1909.09587*
  - [10] Mihaela Bornea, Lin Pan, Sara Rosenthal, Radu Florian, and Avirup Sil. 2020. Multilingual transfer learning for qa using translation as data augmentation. *arXiv preprint arXiv:2012.05958*.
  - [11] Junhao Liu, Linjun Shou, Jian Pei, Ming Gong, Min Yang, and Daxin Jiang. 2020. Cross-lingual machine reading comprehension with language branch knowledge distillation. *arXiv preprint arXiv:2010.14271*
  - [12] Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2018. A span-extraction dataset for chinese machine reading comprehension. *arXiv preprint arXiv:1810.07366*.
  - [13] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2019. Cross-lingual machine reading comprehension. *arXiv preprint arXiv:1909.00361*.
  - [14] Fei Yuan, Linjun Shou, Xuanyu Bai, Ming Gong, Yaobo Liang, Nan Duan, Yan Fu, and Daxin Jiang. 2020. Enhancing answer boundary detection for multilingual machine reading comprehension. *arXiv preprint arXiv:2004.14069*
  - [15] Revanth Gangi Reddy, Md Arafat Sultan, Efsun Sarioglu Kayi, Rong Zhang, Vittorio Castelli, and Avirup Sil. 2020. Answer span correction in machine reading comprehension. *arXiv preprint arXiv:2011.03435*.
  - [16] Zheng, Z. (2017). AnswerBus: Web-based Question Answering System
  - [17] Bandyopadhyay, A. (2019). Dialogue-based Question Answering System for Railway Information.
  - [18] Mikhail Burtsev, Alexander Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nikolay Bushkov, Olga Gureenkova, Taras Khakhulin, Yuri Kuratov, Denis Kuznetsov, Alexey Litinsky, Varvara Logacheva, Alexey Lymar, Valentin Malykh, Maxim Petrov, Vadim Polulyakh, Leonid Pugachev, Alexey Sorokin, Maria Vikhreva, et al.. 2018. DeepPavlov: Open-Source Library for Dialogue Systems. In *Proceedings of ACL 2018, System Demonstrations*, pages 122–127, Melbourne, Australia. Association for Computational Linguistics